Predominance of Methanomicrobiales and diverse hydrocarbon-degrading taxa in the Appalachian coalbed biosphere revealed through metagenomics and genome-resolved metabolisms.

Daniel E. Ross[1,2], Daniel Lipus[1,3,4], and Djuna Gulliver [1*]

[1]National Energy Technology Laboratory, 626 Cochrans Mill Road, P.O. Box 10940, Pittsburgh, PA 15236-0940, USA.

[2]NETL Support Contractor, 626 Cochrans Mill Road, P.O. Box 10940, Pittsburgh, PA 15236-0940, USA

[3]Oak Ridge Institute for Science and Education (ORISE), Oak Ridge, Tennessee.

[4]Section Geomicrobiology, GFZ Geoforschungszentrum Potsdam, Telegrafenberg, Potsdam, Germany

*Corresponding author: djuna.gulliver@netl.doe.gov

Running title: Genome-resolved metabolisms of coalbed methane metagenomes.

Competing interests: The authors declare no competing financial interest.

Originality-Significance Statement.

While the microbiology and effects of hydraulic fracturing of shale plays in the Appalachian Basin has been examined, to date, no study has examined the extant microbial communities of coalbed methane wells in the Appalachian Basin, which represents a microbial reservoir for stimulation of coalbed methane or mitigation of methane release. We combined metagenomics and genome-resolved metabolisms to identify microbial communities and metabolic potential across 8 geographically distinct coalbed methane wells within the Appalachian Basin, revealing the predominance of hydrogenotrophic methanogens from the order Methanomicrobiales, and less abundant, diverse potential hydrocarbon degraders.

Summary

Coalbed deposits are a unique subsurface environment and represent an underutilized resource for methane generation. Microbial communities extant in coalbed deposits are responsible for key subsurface biogeochemical cycling and could be utilized to enhance methane production in areas where existing gas wells have depleted methane stores, or in coalbeds that are unmined, or conversely be utilized for mitigation of methane release. Here we utilize metagenomics and metagenome-assembled genomes to identify extant microbial lineages and genome-resolved microbial metabolisms of coalbed produced water, which has not yet been explored in the Appalachian Basin. Our analyses resulted in the recovery of over 40 metagenome-assembled genomes (MAGs) from eight coalbed methane wells. The most commonly identified taxa among samples were hydrogenotrophic methanogens from the order Methanomicrobiales and these dominant MAGs were highly similar to one another. Conversely, low-abundance coalbed bacterial populations were taxonomically and functionally diverse, mostly belonging to a variety of Proteobacteria classes, and encoding various hydrocarbon solubilization and degradation pathways. The data presented herein provides novel insights into Appalachian Basin coalbed microbial ecology, and our findings provide new perspectives on underrepresented *Methanocalculus* species and low-relative abundance bacterial assemblages in coalbed environments, and their potential roles in stimulation or mitigation of methane release.

Introduction

As conventional natural gas resources become depleted, unconventional gas technologies are emerging with greater importance for global energy security. Coalbed methane (CBM) or coal seam gas is an unconventional gas technology that relies on underutilized natural methane repositories trapped in subsurface coalbeds. CBM wells provide access to subsurface methane deposits at a reduced cost and environmental impact relative to traditional mining practices (Rajaram V, 2005). An estimated 40% of the U.S.-based CBM is biogenic (Thielemann *et al.*, 2004, Strąpoć *et al.*, 2011) and there is growing interest in understanding the biogeochemical processes in coalbed deposits, as these microbial communities may be potential indicators of productive CBM wells and in turn may be utilized to enhance *in situ* production of methane (Ritter *et al.*, 2015). Research in this area offers insights into unique metabolisms of subsurface microbial communities that affect methane production and ultimately, the global energy supply and global carbon cycling.

Biogenic methane production from coal proceeds through three main steps, 1) liberation of fermentable substrates (soluble organic intermediates including alkanes, aromatics, and long-chain fatty acids) from the coal surface, 2) degradation of fermentable substrates, and 3) methanogenesis (Jones *et al.*, 2010, Orem *et al.*, 2010, Bugg *et al.*, 2011). Liberation of fermentable substrates from the coal surface can occur through several coal activation pathways, including methylation, addition to fumarate, hydroxylation, and C1 addition/carboxylation (Tierney & Young, 2010, Widdel & Grundmann, 2010, Widdel & Musat, 2010).

Biological conversion of coal to methane is complex and requires multiple enzymatic steps, performed generally by a diverse set of microorganisms, including hydrocarbon degraders and methanogens (Jones *et al.*, 2010), though recent work has shown a single organism can convert methoxylated aromatic compounds into methane (Mayumi *et al.*, 2016). Coalbeds are complex heterogeneous systems with well to well spatial and temporal variability that shapes coalbed microbial community dynamics (Lawson *et al.*, 2015, Vick *et al.*, 2019). Structurally, coal seams contain a cleat system of connected small pore spaces (<50 nm in diameter), and depending on the coal type and rank, have a range of permeability. The physical inaccessibility and recalcitrance of coal makes the coal-to-methane conversion process slow, and the rate-limiting step is hypothesized to be the initial attack on the coal (Kleinsteuber *et al.*, 2012). Due to the complexity of biological coal-to-methane conversion, the microbial community, and the metabolic pathways responsible for this process, remains largely undefined. Unraveling these pathways and the associated community of microorganisms is paramount to understanding biogenic methane production in Appalachian Basin coal seams, and its contribution to global carbon cycling.

Microbial ecology of subsurface coalbed environments has relied upon 16S rRNA gene sequencing (Strąpoć *et al.*, 2008, Zhang *et al.*, 2015, Barnhart *et al.*, 2016), which provides important information about microbial community structure and diversity but lacks functional data and often yields low taxonomic resolution (genus level and higher) (Hillmann *et al.*, 2018). Furthermore, commonly used PCR primers are subject to amplification bias and can preferentially target specific

taxa (Jones *et al.*, 2015, Eloe-Fadrosh *et al.*, 2016). Metagenomic sequencing can alleviate some of these concerns, providing improved taxonomic resolution and potential metabolic functions of a microbial community, which can be further resolved to genomes. A limited number of studies have utilized metagenomics to characterize the metabolic potential of CBM systems (An *et al.*, 2013, Evans *et al.*, 2015, Lawson *et al.*, 2015, Robbins *et al.*, 2016). To our knowledge, no study has yet explored coalbed microbial assemblages from the Appalachian Basin, which is an energy rich geologic formation that stretches from New York to Alabama and is one of the most productive energy regions in the United States (www.eia.gov).

Here we present an in-depth metagenomic investigation of Appalachian Basin coalbed microbial populations, with correlation to the geochemical signatures of the produced water. Our microbial exploration includes taxonomic profiling of the previously uncharacterized microbial communities using unassembled shotgun metagenome sequencing reads, and genome-resolved functional potential of dominant microbial community members through re-construction and manual curation of metagenome-assembled genomes (MAGs).

Our results demonstrate that the Appalachian Basin is taxonomically distinct from previously characterized coalbed basins, with metagenome samples predominantly comprised of Archaea from the Order Methanomicrobiales. Genome-resolved metabolisms of metagenome-assembled genomes (MAGs) revealed the potential for hydrogenotrophic methanogenesis, as well as saline tolerance mechanisms utilizing compatible solute transport of glycine/betaine. Conversely, bacterial MAGs were less abundant and taxonomically diverse,

including known hydrocarbon degraders (*Pseudomonas, Marinobacter, Hyphomonas,* and *Arcobacter*), sulfate reducers (*Desulfobacteraceae* and *Desulfobulbaceae*), iron reducers (*Shewanella*), and acetogens (*Acetobacterium*). Furthermore, our data suggest that the microbial community composition and potential functionality is driven, in part, by produced water salinity and the available methanogenic substrates. Overall, Appalachian Basin produced water metagenomes contain a taxonomically diverse group of microorganisms responsible for the breakdown of coal substrates into various intermediates that serve as substrates for highly similar hydrogenotrophic methanogens for methane production. Characterization of the extant microbial communities and elucidation of genome-resolved metabolisms in this environment is a necessary first step towards informed strategies for enhanced methane production or mitigation of unwanted methane leakage.

Experimental Procedures.

Experimental workflow is presented in the Supplementary Material (Supplementary Fig. S1).

Sample collection and chemical characterization.

Produced water samples were collected from eight separate coalbed methane wells from the Appalachian Basin. Samples were obtained from wells that were not exposed to any injection fluids or material. The fluid of these samples will be referred to as “produced water” throughout the manuscript. This fluid may have input from meteoric sources, and further research is required to understand the extent of this type of fluid infiltration. Samples were extracted from CBM wells named Key8, Key7, K34, K35, P21, BB137, MC79, and L32A, at a depth of 796 ft., 1,201 ft., 1,704 ft., 1,912 ft., 1,961 ft., 1,980 ft., 2,239 ft., and 2,578 ft., respectively. All samples were obtained from the Pocahontas 3 coal seam in December 2015 [Buchanan County, VA (K34, K35, P21, L32A); Tazewell County, VA (BB137); and McDowell County, WV (MC79)] except for Key8 (Middle Kittanning) and Key7 (Upper Kittanning/Brookville “A”), which came from Indiana County, PA and Westmoreland County, PA, respectively. Produced water samples were collected in sterile 1L bottles, placed on ice (immediately following extraction from the well), and frozen upon arrival at NETL. The coal rank associated with produced water ranged from low/medium volatile bituminous (K34, K35, P21, BB137, MC79, and L32A) to medium/high volatile bituminous (Key7 and Key8). Due to confidentiality agreements with drilling companies, well-specific methane

179 production data was unavailable, therefore the cumulative Appalachian Basin
180 coalbed methane production was estimated between 10 billion cubic feet (BCF) to
181 1 trillion cubic feet (Milici & Polyak, 2014).

182 Major and trace cations and anions (Na, Ca, Mg, K, Fe, Sr, Ba, Li, Mn, and
183 Cl) were measured at the Pittsburgh Analytical Laboratory (National Energy
184 Technology Laboratory, Pittsburgh, Pennsylvania, USA) following EPA methods
185 6010C and 300.1 using ICP-OES (Perkin Elmer Optima 7300 DV) and IC (Dionex
186 Ion Chromatograph). Samples were run in duplicate, with average values being
187 recorded (Supplementary Table S1).

188

189 **DNA extraction, and metagenome sequencing.**

190 Aliquots of coalbed produced water (50 mL) were filtered through a 0.2-µm filter
191 and DNA was extracted from the filter using the MoBio PowerSoil or PowerWater
192 DNA Isolation Kits (Carlsbad, CA), according to manufacturer's instructions. For
193 shotgun metagenome sequencing, DNA libraries were prepared using the Nextera
194 XT DNA Library Preparation Kit according to manufacturer's protocol (Illumina,
195 San Diego, CA). Paired-end sequencing reads (2 x 300 bp) were generated on an
196 Illumina MiSeq with the MiSeq v3 Reagent Kit (600 cycles) (Illumina, San Diego,
197 CA) (Supplementary Table S2).

198

199 **Taxonomy classification of unassembled metagenome reads.**

200 Initial taxonomic inference of reads was performed with Kaiju using the NCBI nr
201 protein database in ‘greedy mode’ with a minimum match length of 11, minimum
202 match score of 75, and 5 allowed mismatches (Menzel *et al.*, 2016). Individual

unassembled metagenome reads from the Appalachian Basin (8 samples, this study), Alberta basin (metagenome accession numbers SRX211003; SRX211004; SRX210875), San Juan Basin (SRX210867; SRX210868; SRX210869; SRX210870), or Powder River Basin (metagenome accession numbers SRX9501106; SRX9501107; SRX9501108) were analyzed with Kaiju. Re-constructed full-length 16S rRNA gene sequences were obtained from metagenome reads using EMIRGE (Miller *et al.*, 2011), within the PhyloFlash software (Gruber-Vodicka *et al.*, 2020).

Read-based microbial abundance correlation.

Read-based taxonomy (Archaea and Bacteria relative abundance) was correlated to major and minor cation/anion concentrations with Spearman rank coefficients using vegan in R (version 2.4-0) (Jari Oksanen, 2013, Team, 2018) (Supplementary Table S1).

Assessment of hydrocarbon degradation potential with unassembled metagenome reads.

Metagenome reads were uploaded to the MG-RAST server. Potential hydrocarbon degradation pathways were examined using COG and subsystems and the total number of hits for specific genes were compiled for each individual metagenome. Hits were normalized to reads per million (RPM), with the total number of reads per metagenome divided by 1,000,000 to obtain a ‘per million scaling factor’. The total number of hits were divided by the per million scaling factor to obtain RPM,

which is represented in Figure 1C. Hydrocarbon degradation pathways examined include oxygen-dependent and oxygen independent activation and dearomatization (An *et al.*, 2013).

Metagenome assembly.

Paired-end reads were assembled individually from each sample using metaSPAdes version 3.8.0 (Nurk *et al.*, 2017) with error correction, and with k-mer values of 33, 55, and 77. Assembly quality metrics (*e.g.,* GC-content, N50, number of contigs, longest contig) were assessed using QUAST (Gurevich *et al.*, 2013) (Supplementary Table S3). Percentage of reads mapping to contigs was performed with the 'map reads to contigs' tool in the CLC Genomics Workbench version 8.5.1 (CLC Bio, Aarhus, Denmark), using the following parameters: no contig masking, match score = 1, mismatch cost = 2, insertion open cost = 6, insertion extend cost = 1, deletion open cost = 6, deletion extend cost = 1, length fraction = 0.5, similarity fraction = 0.8, auto-detect paired distances = on, non-specific match handling = map randomly (Supplementary Table S4).

Metagenome binning, and quality metrics of metagenome-assembled genomes (MAGs).

Contigs generated from individual metagenome assemblies (>500 bp or >1000 bp) were binned using reference-independent methods in the software package VizBin (Laczny *et al.*, 2015) (Supplementary Table S5). Completeness of metagenome-assembled genomes (MAGs) was assessed with CheckM (v1.0.8), using single-

249 copy marker genes (Parks *et al.*, 2015). The lineage_wf was initially employed and
250 bin phylogeny was corroborated with BLASTn (Altschul *et al.*, 1990) and
251 Phylopythia (Gregor *et al.*, 2016). The bins were further interrogated using the
252 user-defined taxonomy_wf in CheckM to obtain genus-level resolution of each
253 MAG, when possible. MAGs generated from contig binning were annotated using
254 the RAST annotation server (version 2.0; RASTtk annotation scheme) (Aziz *et al.*,
255 2008) and PROKKA (Seemann, 2014).

256 Manual curation of MAGs was performed using output from CheckM and
257 RAST. For example, marker genes that were found multiple times throughout the
258 genome were BLASTed against the genome in RAST. The contig containing the
259 'duplicate' marker gene was removed if the length and quality were poor, and the
260 remaining sequence data did not contain any unique protein coding sequences.
261 Quality designation for metagenome assembled genomes (MAGs) was based
262 upon metagenome-assembled genome (MIMAG) (Bowers *et al.*, 2017) standards
263 developed by the Genomic Standards Consortium (GSC), with high-quality draft
264 genomes having >90% completion, <5% contamination, with 23S, 16S, and 5S
265 rRNA genes and >18 tRNAs. Medium-quality draft genomes were characterized
266 by having >50% completion, and <10% contamination, while low-quality draft
267 genomes were <50% complete and <10% contamination (Supplementary Table
268 S5).

269 To further corroborate taxonomic assignment using CheckM, BLASTn, and
270 Phylopythia, the most complete 20 MAGs (>50% complete, <10% contamination)
271 were uploaded to the KBase server (www.kbase.us) and taxonomic assignments

for MAGs were examined with the 'Classify Microbes with GTDB-Tk' (V1.7.0) (Chaumeil *et al.*, 2019) (Supplementary Table S5). Initial metabolic characterization of each individual MAG was performed with the 'Annotate and Distill Assemblies with DRAM' application (kb_v0.1.0) with a minimum contig length of 1000, bit score threshold of 60, and reverse bit score threshold of 350. Within DRAM, Prodigal (Hyatt *et al.*, 2010) was used to detect open reading frames (ORFs) and predict amino acid (AA) sequences. Characterization (annotation) of AA sequences was performed using the KEGG, UniRef90, and MEROPS databases (Shaffer *et al.*, 2020). Specific hits to CRISPR sequences were identified for each MAG and compiled into a single file, including gene description, KEGG number, and CRISPR type/subtype (Supplementary File 3). For CRISPR-Cas proteins, the total number of CRISPR hits for each genome/MAG was normalized to genome/MAG size by dividing the total number of CRISPR-Cas hits by the genome/MAG size (in bp) and multiplying by a million. CRISPR repeats and CRISPR spacers were identified in RAST (Supplementary File 3).

Phylogenetic analysis using 16S rRNA gene sequences from *Methanocalculus* MAGs.

The 16S rRNA gene sequence was identified in each *Methanocalculus* MAG and sequences were compiled with 16S rRNA gene sequences from top BLASTn hits using MEGA (Tamura *et al.*, 2013). Briefly, 16S rRNA gene sequences from *Methanocalculus* MAGs were individually BLASTed against the NCBI nr/nt nucleotide collection, optimized for highly similar sequences (megablast). The best

matches to each 16S rRNA gene sequence, including cultured and uncultured representative sequences, were used to construct a phylogenetic tree using the Maximum Likelihood method (Tamura-Nei model). Details on tree construction are provided in the Supplementary Figure S5 legend.

Phylogenetic analysis of Methanomicrobiales MAGs using McrA.

McrA protein sequences were identified through manual examination of genomes in RAST (Aziz *et al.*, 2008). McrA sequences from the *Methanocalculus* and *Methanoregula* MAGs were analyzed using the non-redundant (nr) protein database in the BLAST algorithm of the NCBI database. Top hits (query coverage =100%, percent identity >80%, E-value = 0) from Methanomicrobiales were combined into a single fasta file with the CBM McrA sequences. Sequences were aligned using MUSCLE (Edgar, 2004), and a Neighbor-Joining tree (bootstrap value = 1000) was generated using MEGA using default settings (Tamura *et al.*, 2013). Details on tree construction are provided in the Supplementary Figure S6 legend.

Average nucleotide identity/average amino acid identity of Methanomicrobiales MAGs.

Nucleotide sequences and amino acid sequences for each MAG were uploaded to the enve-omics ANI/AAI distance matrix calculator to calculate an all against all average nucleotide identity and average amino acid identity across all MAGs

317 (http://enve-omics.ce.gatech.edu/g-matrix/) (Goris *et al.*, 2007, Rodriguez-R &
318 Konstantinidis, 2016) (Supplementary Table S6).
319
320 **Data availability.**
321 The datasets generated and/or analyzed during the current study are available on
322 the Energy Data Exchange (EDX) page
323 (https://edx.netl.doe.gov/dataset/appalachian-basin-metagenome-sequencing-
324 reads). Appalachian Basin metagenome reads were deposited under the
325 BioProject PRJNA874672. Metagenome read files and metagenome-assembled
326 genomes can also be accessed via the kBase webserver
327 (https://narrative.kbase.us/narrative/105685). Sequence data for the Alberta Basin
328 (CO182, CO183, Trident 1560D) and San Juan Basin (CG7, CG8, CG13, CG19)
329 were retrieved from the SRA database under the BioProject PRJNA183510.
330 Powder River Basin samples were retrieved from the SRA database under the
331 BioProject PRJNA678021.
332
333

Results

Appalachian Basin produced water geochemistry.

Produced water was collected from eight Appalachian Basin coalbed methane wells, ranging in depth from 796 ft to 2,578 ft (Fig. 1). Samples were moderately to highly saline, with major ions (*e.g.,* Na, Cl, Mg, and Ca) ranging in concentration from 7.29 g/L (sample Key8) to 130.3 g/L (sample K34) (Fig. 1; Supplementary Table S1). Concentrations of barium, strontium, and lithium were elevated in samples with the highest concentration of major ions (*e.g.*, samples BB137 and K34). Appalachian Basin produced waters contained primarily sodium and chloride, with lower concentrations of magnesium and sulfate (Supplementary Table S1; Supplementary Fig. S2 and S3), a common geochemical signature of produced waters associated with coalbed methane (CBM) (Van Voast, 2003). Overall, sample depth did not correlate with salinity, as some shallower samples had higher salinity compared to deeper samples (Fig. 1).

Taxonomic identification of Appalachian Basin coal seam produced water metagenomes and comparison to coal seam metagenomes from geographically distinct coal basins.

A total of 1.6-4.4 million reads were generated for each sample using high-throughput sequencing of metagenome libraries, with a total of ~351-745 Mbp per sample (Supplementary Table S2). An initial estimation of microbial community composition of Appalachian Basin metagenomes was determined using reconstructed full-length 16S rRNA gene sequences from metagenome

sequencing reads (Supplementary File 1). The most prevalent taxa included *Methanocalculus halotolerans*, *Methanoregula formicica*, and *Methanobacterium* spp., including *M. flexile, M. formicicum, M. movans,* and *M. paludis*. Sequences related to *Pseudomonas* were found in three metagenome datasets, two most closely related to *Pseudomonas stutzeri*, and one most closely related to *Pseudomonas fluorescens*. *Desulfovibrio* was observed in three metagenome datasets, two most closely related to *Desulfovibrio aespoensis*, and one most closely related to *Desulfovibrio vulgaris* (Supplementary File 1).

Expanding upon 16S rRNA gene taxonomy, we utilized metagenome sequencing reads to determine relative abundances of microorganisms in each sample. Taxonomic profiling of metagenome reads revealed that Appalachian Basin metagenomes contained varied proportions of both Archaea and Bacteria, with the relative abundance of Archaea >48% in 5 out of 8 samples (Fig. 2). Within the domain Archaea three major groups of methanogens were detected, including known acetoclastic and methylotrophic methanogens from the orders Methanosarcinales and Methanobacteriales, and hydrogenotrophic methanogens from the order Methanomicrobiales. Members of the order Methanomicrobiales had the highest representation in most of the samples, with relative abundances reaching 69% (Fig. 2). The most abundant genus of this group was Methanocalculus (Ollivier *et al.*, 1998, Mori *et al.*, 2000, Lai *et al.*, 2002, Lai *et al.*, 2004, Zhilina *et al.*, 2013, Sorokin *et al.*, 2015).

In samples where bacterial sequences were dominant (>50% relative abundance), Proteobacteria had the highest representation. The most abundant

classes were Gammaproteobacteria, Alphaproteobacteria, and Deltaproteobacteria (Fig. 2). Similar bacterial taxa have previously been detected in other coalbed basins (Strąpoć *et al.*, 2008, Zhang *et al.*, 2015, Barnhart *et al.*, 2016), and crude oil reservoirs (Shelton *et al.*, 2016). Conversely, metagenomes containing the lowest relative abundances of Bacteria (<40%) had a higher proportion of members from the Terrabacteria group (Firmicutes, Chloroflexi, and Actinobacteria), FCB group (Bacteroidetes), and Synergistales.

Next, we compared the Appalachian Basin (AppB) metagenome taxonomy to metagenome datasets from three geographically distinct coal basins, the Alberta Basin (AlbB) (Alberta, Canada), San Juan Basin (SJB) (Colorado, USA), and Powder River Basin (PRB) (Montana, USA) (An *et al.*, 2013, Schweitzer *et al.*, 2022). The Appalachian Basin metagenomes (unassembled reads) were taxonomically distinct from the Alberta Basin, San Juan Basin, and Powder River Basin metagenomes (Fig. 2). The relative abundance of Archaea in Appalachian Basin metagenomes ranged from 1% (sample Key8) to 79% (sample BB137), with five of the eight samples containing >48%. In contrast, metagenomes from the Alberta Basin, San Juan Basin, and Powder River Basin had Archaea relative abundances ranging from 0-10%, with the majority containing 1-2%. Most notably, Methanomicrobiales were dominant in the majority of Appalachian Basin metagenomic datasets (41-69% relative abundance), while the Alberta Basin, San Juan Basin, and Powder River Basin metagenomic datasets contained minor amounts of Methanomicrobiales (0.1-3% relative abundance) (Fig. 2). Other notable differences between samples included a higher relative proportion of

403 Methanocella (sample CG19, SJB) Desulfobacterales (Key8, AppB; Terret and
404 Nance, PRB), Chloroflexi (all PRB samples), Selenomonadales (CG8, SJB),
405 Desulfovibrionales (CG8, SJB), Thermoanaerobacteriales (CG8 and CG19, SJB),
406 Campylobacterales (CG13, SJB), Burkholdariales (CO182 and CO183, AlbB),
407 Rhodocyclales (CO182 and CO183, AlbB; CG19, SJB), Rhizobiales, (Nance and
408 Flowers-Goodale, PRB) and Rhodobacterales (Trident1560, AlbB).
409
410 *Correlation of taxonomy and geochemistry.*
411 Following geochemical characterization of produced water and taxonomic
412 identification of extant microbial communities, we evaluated potential links
413 between produced water geochemistry and microbial community composition.
414 This analysis revealed a statistically significant positive correlation between the
415 relative abundance of Archaea and salinity (Fig. 1, Supplementary Table S1).
416 Conversely, a negative correlation was observed between the relative abundance
417 of Bacteria in metagenome samples and salinity (Fig. 1, Supplementary Table S1).
418 Specifically, sodium correlated strongly with the relative abundance of Archaea
419 (R=0.91; P=0.002), and negatively with the relative abundance of Bacteria (R=-
420 0.91; P=0.002). Statistical analysis also suggested a moderate correlation
421 between Archaea and Mg, Ca, or Cl concentration (R=0.86, 0.81, and 0.86,
422 respectively). Other cations/anions including Mn, Ni, Co, Zn, Sr, Ba, and Li
423 followed this trend as well.
424
425 *Characterization of metagenome-assembled genomes (MAGs).*

We assembled high-quality paired-end metagenome reads, which yielded ~24,000-149,000 contigs, with a total assembly size of 24.3-109 Mbp (Supplementary Table S3; Supplementary Figure S4). To assess the quality of metagenome assembly, metagenome reads were mapped to the assembled metagenome contigs. The percentage of reads mapping to the metagenome assembly ranged from 87-94%. Next, contigs from metagenome assemblies were binned, and metagenome-assembled genomes (MAGs) were reconstructed. Metagenome reads were then mapped to each individual MAG and ranged from 77-94% (Supplementary Table S4). Metagenome contig binning resulted in the recovery of 40 MAGs (21 medium- to high-quality MAGs), across 8 metagenomes (Supplementary Table S5). Archaeal MAGs were represented by a single order (Methanomicrobiales), while bacterial MAGs spanned five taxa, including Pseudomonadales, Alteromonadales, Campylobacterales, Desulfobacterales, Clostridiales, and Rhodobacterales.

Archaeal MAGs

Metagenome assembly and binning produced ten Methanomicrobiales metagenome-assembled genomes (MAGs) from seven different samples. Six were most closely related to *Methanocalculus* and four were most closely related to *Methanoregula*. The *Methanocalculus* MAGs ranged in size from 2.0 to 2.9 Mbp and had relative abundances ranging from 5.21% to 68.0% (Supplementary Table S4). Genome coverage of the six *Methanocalculus* MAGs (total bases mapped/total MAG assembly length) ranged from 16.8x to 333x, and order-level

449 completeness (Methanomicrobiales) was >97% with <5% contamination
450 (Supplementary Table S4 and S5). Based upon 16S rRNA gene recovered in the
451 MAGs, six of the Methanomicrobiales MAGs were most closely related to
452 *Methanocalculus halotolerans* SEBR 4845, which originated from a low-
453 temperature biodegraded oil reservoir (Ollivier *et al.*, 1998) (Supplementary Fig.
454 S5).

455 The four *Methanoregula* MAGs had relative abundances ranging from 5.2
456 to 15.8%. Genome sizes ranged from 2.8-2.9 Mbp, with a genome coverage of 16
457 to 91x (Supplementary Table S4 and S5). Based on the 16S rRNA gene recovered
458 from three of the four MAGs, the *Methanoregula* MAGs were most closely related
459 to *Methanoregula formicica* (CP003167), which was isolated from a propionate-
460 degrading enrichment culture of an anaerobic sludge blanket reactor and utilizes
461 H_2/formate as substrates for methanogenesis (Yashiro *et al.*, 2011, Yamamoto *et*
462 *al.*, 2014).

463 Next, we examined the *mcrA* gene that encodes the methyl coenzyme-M
464 reductase (Mcr), which performs the terminal step in methanogenesis and is a
465 robust alternative to 16S rRNA gene analysis for identification and phylogenetic
466 placement of methanogens (Luton *et al.*, 2002). Identification of full-length McrA
467 amino acid sequences and subsequent phylogenetic analysis with closely related
468 methanogens revealed identical placement of the 10 Methanomicrobiales MAGs
469 into *Methanocalculus* or *Methanoregula* clades (Supplementary Fig. S6), similar to
470 16S rRNA gene sequence analysis. McrA protein sequences were highly

conserved within the *Methanocalculus* or *Methanoregula* MAGs with 97-98% identity and 99% positives for the full-length protein.

The tight branching of the *Methanocalculus* or *Methanoregula* MAGs suggests that within their respective clades these MAGs are potentially members of the same species. To further examine their relatedness, whole genome comparisons using average nucleotide identity (ANI) and average amino acid identity (AAI) was employed. The Methanomicrobiales MAGs were highly similar across the Appalachian Basin, with ANI and AAI values >95%, suggesting that five of the six *Methanocalculus* MAGs belong to the same species, and three of the four *Methanoregula* MAGs belong to the same species (Supplementary Table S6). Furthermore, the *Methanocalculus* MAGs were unique compared to other Methanomicrobiales MAGs from hydrocarbon resource environments (ANI = 86-92%), and more distant with other sequenced Methanomicrobiales (Supplementary Table S6; Supplementary Fig. S7).

Bacterial MAGs

Bacterial MAGs were reconstructed from most of the metagenomes and included Gammaproteobacteria (*Pseudomonas, Marinobacter,* and *Shewanella*), Alphaproteobacteria (*Hyphomonas* and *Roseovarius*), Deltaproteobacteria (*Desulfosarcina,* and *Desulfocapsa*), and Epsilonproteobacteria (*Acrobacter*). Based upon metagenome read recruitment to each MAG, the *Pseudomonas* MAGs had the highest relative abundance (61%), followed by *Desulfosarcina* (36%), *Shewanella* (18%), *Desulfocapsa* (8.9%), *Hyphomonas* (4.3%),

494 *Roseovarius* (2%), and *Marinobacter* (1.9%) (Supplementary Table S4). The
495 Bacterial MAGs varied in quality, with completeness values ranging from 99%
496 (*Pseudomonas* K35) to 58% (*Roseovarius* L32A) (Supplementary Table S5).
497 Next, we examined the taxonomic placement of each MAG using 49
498 universal marker genes. MAGs were grouped with the closest sequenced
499 genomes, which included *Acetobacterium, Desulfocapsa, Pseudomonas,*
500 *Marinobacter, Hyphomonas,* and *Arcobacter* (Supplementary Fig. S8). The
501 *Roseovarius* L32A MAG, *Desulfosarcina* Key8 MAG, and the *Geobacter* K34 MAG
502 did not have any close relatives.
503
504 Methanocalculus *and* Methanoregula *shape coalbed Archaea populations and are*
505 *likely primary contributors to methanogenic activity.*
506 To ascertain the potential role of the Methanomicrobiales MAGs in biogenic
507 methane production in Appalachian Basin coalbeds, we examined
508 methanogenesis pathways. Methanogenic archaea can utilize several electron
509 donors for CO_2 reduction, including H_2 and formate (hydrogenotrophic), acetate
510 (acetoclastic), methanol and methyl amines (methylotrophic), and other alternative
511 substrates, including ethanol, CO, choline, and betaine (Kurth *et al.*, 2020).
512 Metabolic reconstruction of the *Methanocalculus* and *Methanoregula* MAGs
513 revealed the presence of genes associated with archaeal methane metabolism
514 (Fig. 3; Supplementary File 2), specifically hydrogenotrophic methanogenesis.
515 Encoded pathways included the methyl-coenzyme reductase (MCR) complex and
516 Na^+-translocating methyl-H4MPT:coenzyme M methyltransferase

517 (*mtrABCDEFG*). Energy-conserving complexes were present and included
518 energy-converting hydrogenase (*eha*) (*Methanocalculus*), energy-conserving
519 hydrogenase (*echA-F*) (*Methanoregula*), coenzyme F_{420} hydrogenase (*frhABG*),
520 and V-type ATP synthase. All Methanomicrobiales MAGs contain genes encoding
521 for the CoB--CoM heterodisulfide reductase complex (*hdrABC*), which is
522 responsible for cycling of coenzyme M (CoM) and coenzyme B (CoB). The
523 absence of genes for methyl compound utilization [methanol (*mtbA*), methylated
524 amines (*mtaA, mttBC, mtbBC*), and methyl sulfides (*mtsA*)], suggests
525 *Methanocalculus* and *Methanoregula* are unable to utilize these compounds as
526 methanogenic substrates and are most likely hydrogenotrophic methanogens (Fig.
527 3; Supplementary File 2).
528
529 *Potential for hydrocarbon degradation in Appalachian Basin metagenomes and*
530 *metagenome-assembled genomes (MAGs).*
531 The conversion of coal into methane requires a variety of microorganisms with
532 varied metabolisms (Vick *et al.*, 2019, Vick *et al.*, 2019). The initial attack on the
533 coal to liberate fermentable substrates is considered the rate-limiting step, which
534 is achieved by hydrocarbon-activating (degrading) bacteria and methanogens
535 (Strąpoć *et al.*, 2008, Jones *et al.*, 2010, Kleinsteuber *et al.*, 2012). Polycyclic
536 aromatic hydrocarbons (PAHs) are a major constituent in coalbed production water
537 and PAH-degrading microorganisms are prevalent in coal reservoirs (Penner *et*
538 *al.*, 2010, Meslé *et al.*, 2013). Hydrocarbon degradation pathways were present in
539 all metagenomes, though samples containing a higher relative abundance of

Bacteria sequences had a higher normalized proportion of hydrocarbon degradation pathways (Fig. 1). Specifically, many but not all metagenomes encoded for toluene degradation (toluene→benzoate), biphenyl degradation (biphenyl→benzoate), benzoate degradation (benzoate→catechol), and naphthalene degradation (naphthalene→catechol) (Fig. 4; Supplementary Fig. S9-S11). Pathways for conversion of catechol to acetyl-CoA/succinyl-CoA (catechol ortho-cleavage) or acetyl-CoA (catechol meta-cleavage) were present in many of the metagenome samples (Fig. 4; Supplementary Fig. S10 and S12). The anaerobic conversion of toluene to benzoyl-CoA (addition of fumarate) was encoded in four of the metagenomes (Key8, Key7, P21, and BB137) (Fig. 4; Supplementary Fig. S13).

Detailed examination of Appalachian Basin MAGs revealed the potential for hydrocarbon degradation. Specifically, *Hyphomonas* and *Marinobacter* encode for phenol hydroxylase, with the potential to convert benzene to catechol. *Marinobacter, Hyphomonas,* and *Roseovarius* encode for the ability to convert toluene to benzoate, while the three *Pseudomonas* MAGs encode for benzoate 1,2-dioxygenase for conversion of benzoate to methylcatechol (Fig. 4; Supplementary File 2). Pathways for the degradation of benzoate, biphenyl, and catechol, as well as n-alkane hydroxylation and beta oxidation were encoded by the *Roseovarius* MAG. Furthermore, catechol, phenylacetate, and 2-oxopentanoate degradation, as well as the ortho-cleavage pathway, were encoded by the *Hyphomonas* MAG (Fig. 4; Supplementary File 2).

563 *Factors influencing microbial survival and persistence in Appalachian Basin*
564 *coalbeds.*

565 Members of the *Methanocalculus* genus have previously been found to
566 tolerate up to 12% NaCl (120 g/L) (Ollivier *et al.*, 1998). The Appalachian Basin
567 *Methanocalculus* MAGs were recovered from produced water containing salinities
568 up to 130.3 g/L, therefore we examined potential pathways for saline tolerance.
569 Saline tolerance is achieved through a variety of mechanisms, involving salt
570 accumulation (K^+ or Cl^-) (*i.e.,* salt-in strategy), or accumulation/synthesis of organic
571 solutes, such as glycerol, glycine-betaine, ectoine, and trehalose (Oren, 2008).
572 The *Methanocalculus* MAGs encoded for genes responsible for compatible solute
573 transport [glycine/betaine (*opu*)] but lacked genes for compatible solute production
574 and salt-in strategies (Supplementary File 2). Pathways for saline tolerance in
575 *Methanoregula* were less defined, with three of the four MAGs encoding a few
576 subunits of the glycine/betaine ABC transport system (*otaB, otaC,* and *opuAC*) and
577 a sodium/glycine symporter (*glyP*).

578 Although a negative correlation between bacterial relative abundance and
579 anion/cation concentration was evident, saline tolerance pathways utilizing the
580 salt-in strategy or proline biosynthesis were present in the Bacteria MAGs. Salt-in
581 strategies include 2-component system OmpR family sensor histidine kinase
582 (*Pseudomonas*, *Acetobacterium*, *Hyphomonas*), K+ transporting ATPase
583 (*Pseudomonas*, *Acetobacterium*, *Hyphomonas*), Na+:H+ antiporter
584 (*Pseudomonas*, *Acetobacterium*, *Desulfosarcina*, *Desulfocapsa, Hyphomonas,*
585 *Arcobacter, Marinobacter, Shewanella*), and KtrAB potassium uptake system

586 (*Pseudomonas, Acetobacterium, Arcobacter, Shewanella*). Furthermore,
587 *Pseudomonas* MAGs potentially utilize compatible solute production, as pathways
588 for trehalose biosynthesis and ectoine biosynthesis were present (Supplementary
589 File 2).

590 Viral immunity through CRISPR-Cas systems contributes to microbial
591 persistence in subsurface environments (Daly *et al.*, 2016). Examination of our
592 Appalachian Basin MAGs revealed the presence of CRISPR-Cas systems in 13 of
593 the 21 highest quality MAGs (Supplementary Table S7). *Methanocalculus* MAGs
594 had the greatest number of CRISPR-Cas proteins (17-48), followed by the
595 *Desulfocapsa* Key8 MAG (19), the *Desulfosarcina* Key8 MAG (18), and the
596 *Arcobacter* L32A MAG (11). The remaining MAGs contained 2 or less CRISPR-
597 Cas proteins. Specifically, the *Methanocalculus* MAGs encoded for various
598 CRISPR-Cas subtypes, including Subtype 1-A, Subtype 1-C, Subtype 1-D,
599 Subtype 1-E, Subtype II-B, Subtype III-A, Subtype III-B, and universal Cas proteins
600 Cas1 and Cas2. Of the MAGs encoding for CRISPR-Cas proteins, the number of
601 CRISPR repeats and CRISPR spacers for all Appalachian Basin MAGs was
602 varied, ranging from 16-157 repeats and 14-150 spacers (Supplementary File 3).

603 To determine if the higher number of CRISRP-Cas proteins was due to a
604 common feature of the *Methanocalculus* genus or a specific adaptation to our
605 coalbed strains, we examined nine *Methanocalculus* MAGs from various
606 environments alongside the six Appalachian Basin *Methanocalculus* MAGs. For
607 each MAG, the total number of sequences annotated as CRISPR-Cas sequences
608 was normalized to genome size (number of CRISPR-Cas sequences/genome

609 size*1,000,000). The Appalachian Basin *Methanocalculus* MAGs had a total
610 number of normalized CRISPR proteins ranging from 8.4 to 17.7. Similarly, MAGs
611 that originated from an oil reservoir (*Methanocalculus* 52_23), or an anoxic soda
612 lake sediment (*Methanocalculus* MSAOArc2) contained the most normalized
613 CRISPR-Cas proteins (17.7 and 24.3, respectively), while others from anaerobic
614 digesters or sedimentary rock encoded for fewer normalized CRISPR-Cas proteins
615 (0-6.6).
616

617 **Discussion**

618 We sampled coalbed produced water from the Appalachian Basin, which included
619 eight geographically distinct CBM wells in Southwestern Pennsylvania, West
620 Virginia, and Virginia. Samples originated from wells with various depths (796 feet
621 to 2,578 feet) and contained a wide range of salinities (7.6 g/L to 130.3 g/L). We
622 found a strong correlation between Archaeal sequences and salinity, and a
623 negative correlation between Bacterial sequences and salinity. We recovered over
624 40 metagenome-assembled genomes (MAGs), with most samples containing
625 MAGs from the order Methanomicrobiales. Within the Methanomicrobiales, five
626 MAGs from the *Methanocalculus* genus and three from the *Methanoregula* genus
627 exhibited species-level overlap (ANI >95%). In contrast, the bacterial populations
628 were taxonomically more diverse than the Archaea population, with MAGs related
629 to *Pseudomonas, Desulfocapsa, Desulfosarcina, Hyphomonas, Roseovarius,*
630 *Arcobacter, Marinobacter,* and *Shewanella.*

631 Methanomicrobiales are hydrogenotrophic methanogens inhabiting a
632 variety of environments, including marine and freshwater sediments (Liu &
633 Whitman, 2008), deep sand aquifers (Katayama *et al.*, 2015), oil sands reservoirs
634 (Hubert *et al.*, 2012) and coalbed environments across the United States and
635 abroad (Strąpoć *et al.*, 2008, Gallagher *et al.*, 2013, Zhang *et al.*, 2015, Beckmann
636 *et al.*, 2019). *Methanocalculus* spp. dominated most of the Appalachian Basin
637 coalbed methane archaeal community and while our MAGs were distinct from
638 other sequenced *Methanocalculus* spp., this genus has been observed in the
639 planktonic communities of coal microcosms from the Sydney Basin (Vick *et al.*,
640 2019, Vick *et al.*, 2019), microbial communities from Indonesian coals (Susilawati
641 *et al.*, 2015), produced water samples from the San Juan Basin (An *et al.*, 2013),
642 and Powder River Basin metagenomes (Schweitzer *et al.*, 2019).
643 Common coalbed community functions (*e.g.,* hydrocarbon degradation) are
644 encoded by a diverse set of microorganisms, with functional redundancies across
645 taxa (Vick *et al.*, 2019). Specifically, microorganisms from coal enrichment cultures
646 from the Bowen, Surat, Sydney coal basins of eastern Australia, including
647 *Chelatococcus, Citrobacter, Pseudomonas, Marinobacter,* and *Thauera* spp.
648 encode for aromatic hydrocarbon degradation genes/pathways (*e.g.*, benzoate,
649 benzene, toluene) (Vick *et al.*, 2019). These communities were markedly different
650 from hydrocarbon degrading communities of the Powder River Basin, (*Chlorflexi,*
651 *Bacteroidetes,* and *Geobacter*) (McKay *et al.*, 2022), and the Appalachian Basin
652 (*Pseudomonas, Marinobacter, Hyphomonas,* and *Roseovarius*), though a few
653 similar taxa are found across coal basins (*e.g., Pseudomonas* and *Marinobacter*).

Environmental factors, such as salinity, coal rank (maturity), and depth, are known to play a key role in shaping the taxonomic distribution of subsurface microbial communities (Lozupone & Knight, 2007, Morrissey & Franklin, 2015, Hoehler *et al.*, 2018). In the Appalachian Basin, the persistence of *Methanocalculus* and accompanying bacterial taxa may be influenced by several factors, namely substrate availability, salinity, and viral predation. Due to the inherent carbon complexity of coal seam environments, the type and variety of available fermentable substrates will drive intermediate formation, and thus contribute to the emergence of various taxa (Vick *et al.*, 2019). Specifically, Vick and coworkers observed a successional pattern of *Methanocalculus* abundance, where abundances increased rapidly over time and then abruptly dropped. It was hypothesized that a limited resource or intermediate was utilized as it was slowly liberated from the coal surface (Vick *et al.*, 2019). For example, in the Powder River Basin, acetate is a key intermediate for biological conversion of coal to methane (McKay *et al.*, 2022). Microbial communities from the Appalachian Basin were markedly different than the Powder River Basin and no genomic evidence was found to support acetoclastic methanogensis, suggesting an alternative mechanism is employed. Based upon our findings, we hypothesize that the dominance and genomic conservation of *Methanocalculus* MAGs and the diversity of the bacterial population in Appalachian Basin samples, is driven, in part, by the available methanogenic substrates at the time of sampling (Waldron *et al.*, 2007).

Methanogenesis is influenced by salinity in coal environments (Kirk *et al.*, 2015), and we found a positive correlation between Archaea relative abundance

and salinity and a negative correlation between Bacteria relative abundance and salinity in Appalachian Basin coal seam samples (Supplementary Table S1). It is unknown whether this correlation is due to a) the saline tolerance mechanisms employed in Archaea versus Bacteria, or b) a correlation between salinity and carbon availability. Salinity tolerance mechanisms were found in both Archaeal MAGs and Bacterial MAGs. However, the mechanisms by which this is achieved varied between the two groups. For example, the Methanomicrobiales MAGs encoded for pathways specific to compatible solute transport, specifically glycine/betaine transport and utilization, while the Bacterial MAGs likely employ a salt-in strategy or rely on biosynthesis of proline (Fichman *et al.*, 2015). Carbon substrate content was not measured in this study, however, hydrocarbon degradation potential in saline environments decreases due to the salting out effect. Namely, the solubility of hydrocarbons decreases as salinity increases, and in the case of coal seams, becomes adsorbed to the coal and subsequently less bioavailable (Margesin & Schinner, 2001). This may have resulted in a secondary effect, in which the salinity was an indicator of carbon availability, and carbon availability was the actual driver for the microbial community composition.

Similar to other subsurface environments we found evidence of CRISPR-Cas systems, primarily Type 1 CRISPR-Cas machinery, which are responsible for viral immunity (Daly *et al.*, 2016). Specifically, CRISPR-Cas machinery was found in all *Methanocalculus* MAGs, while only three bacterial MAGs encoded for a CRISPR-Cas system (*Arcobacter, Desulfocapsa,* and *Desulfosarcina*) (Supplementary File 3). Daly et al. suggest a high viral predation drove adaptability

in the subsurface environment that selected for resilient microorganisms (Daly *et al.*, 2016). Interestingly, Daly et al. also found a predominance of primarily Type 1 of the CRISPR-Cas machinery that we find in our system. A similar mechanism may be operational for microbial persistence in Appalachian Basin coal seams, with *Methanocalculus* MAGs harboring the most CRISPR-Cas proteins, spacers, and repeats, though further work is needed to understand how viral defense correlates to microbial persistence in subsurface coal seams.

Conclusions.

Taxonomic characterization of Appalachian Basin coalbed produced water metagenomes, combined with genome-resolved microbial metabolisms of subsurface coalbed systems, revealed abundant populations (Methanomicrobiales) that were highly similar (low diversity) and ubiquitous within the Appalachian Basin, suggesting biogenic methane production proceeds mainly through hydrogenotrophic methanogenesis pathways. Appalachian Basin coal seam microbial communities were comprised of distinct microbiological assemblages compared to other coal basins, and our findings suggest novel microbial assemblages are driven, in part, by salinity and the complexity and availability of carbon substrates. The data presented herein provides novel insights into Appalachian Basin microbial ecology, and our findings provide new perspectives into potential pathways for biogenic methane production in Appalachian Basin coalbeds.

Disclaimer:

This work was funded by the Department of Energy, National Energy Technology Laboratory, an agency of the United States Government, through a support contract with Leidos Research Support Team (LRST). Neither the United States Government nor any agency thereof, nor any of their employees, nor LRST, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Author Contributions.

D.E.R. performed experiments, analyzed data, and wrote the manuscript; D.L. performed experiments, analyzed data, and edited the manuscript; D.G. analyzed data and wrote the manuscript.

Declaration of competing interest:

The authors declare no competing financial interest.

References

Altschul SF, Gish W, Miller W, Myers EW & Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215**: 403-410.

An D, Caffrey SM, Soh J*, et al.* (2013) Metagenomics of hydrocarbon resource environments indicates aerobic taxa and genes to be unexpectedly common. *Environmental science & technology* **47**: 10708-10717.

Aziz RK, Bartels D, Best AA*, et al.* (2008) The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics* **9**: 75.

Barnhart EP, Weeks EP, Jones EJP*, et al.* (2016) Hydrogeochemistry and coal-associated bacterial populations from a methanogenic coal bed. *International Journal of Coal Geology* **162**: 14-26.

Beckmann S, Luk AWS, Gutierrez-Zamora M-L, Chong NHH, Thomas T, Lee M & Manefield M (2019) Long-term succession in a coal seam microbiome during in situ biostimulation of coalbed-methane generation. *The ISME Journal* **13**: 632-650.

Bowers RM, Kyrpides NC, Stepanauskas R*, et al.* (2017) Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature Biotechnology* **35**: 725-731.

Bugg TD, Ahmad M, Hardiman EM & Rahmanpour R (2011) Pathways for degradation of lignin in bacteria and fungi. *Nat Prod Rep* **28**: 1883-1896.

Chaumeil P-A, Mussig AJ, Hugenholtz P & Parks DH (2019) GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**: 1925-1927.

Daly RA, Borton MA, Wilkins MJ*, et al.* (2016) Microbial metabolisms in a 2.5-km-deep ecosystem created by hydraulic fracturing in shales. *Nature Microbiology* **1**: 16146.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**: 1792-1797.

Eloe-Fadrosh EA, Ivanova NN, Woyke T & Kyrpides NC (2016) Metagenomics uncovers gaps in amplicon-based detection of microbial diversity. *Nature Microbiology* **1**: 15032.

Evans PN, Parks DH, Chadwick GL, Robbins SJ, Orphan VJ, Golding SD & Tyson GW (2015) Methane metabolism in the archaeal phylum Bathyarchaeota revealed by genome-centric metagenomics. *Science* **350**: 434-438.

Fichman Y, Gerdes SY, Kovács H, Szabados L, Zilberstein A & Csonka LN (2015) Evolution of proline biosynthesis: enzymology, bioinformatics, genetics, and transcriptional regulation. *Biological Reviews* **90**: 1065-1099.

Gallagher LK, Glossner AW, Landkamer LL, Figueroa LA, Mandernack KW & Munakata-Marr J (2013) The effect of coal oxidation on methane production and microbial community structure in Powder River Basin coal. *International Journal of Coal Geology* **115**: 71-78.

Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P & Tiedje JM (2007) DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol* **57**: 81-91.

Gregor I, Dröge J, Schirmer M, Quince C & McHardy AC (2016) PhyloPythiaS+: a self-training method for the rapid reconstruction of low-ranking taxonomic bins from metagenomes. *PeerJ* **4**: e1603.
Gruber-Vodicka HR, Seah BKB, Pruesse E, Arumugam M & Kato S (2020) phyloFlash: Rapid Small-Subunit rRNA Profiling and Targeted Assembly from Metagenomes. *mSystems* **5**: e00920-00920.
Gurevich A, Saveliev V, Vyahhi N & Tesler G (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**: 1072-1075.
Hillmann B, Al-Ghalith Gabriel A, Shields-Cutler Robin R, Zhu Q, Gohl Daryl M, Beckman Kenneth B, Knight R, Knights D & Rawls John F (2018) Evaluating the Information Content of Shallow Shotgun Metagenomics. *mSystems* **3**: e00069-00018.
Hoehler T, Losey NA, Gunsalus RP & McInerney MJ (2018) Environmental Constraints that Limit Methanogenesis. *Biogenesis of Hydrocarbons,*(Stams AJM & Sousa D, eds.), p.^pp. 1-26. Springer International Publishing, Cham.
Hubert CRJ, Oldenburg TBP, Fustic M*, et al.* (2012) Massive dominance of Epsilonproteobacteria in formation waters from a Canadian oil sands reservoir containing severely biodegraded oil. *Environmental Microbiology* **14**: 387-404.
Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW & Hauser LJ (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**: 119.
Jari Oksanen FGB, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, R.B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs, Helene Wagner (2013) Vegan: Community Ecology Package. *CRAN*.
Jones EJP, Voytek MA, Corum MD & Orem WH (2010) Stimulation of methane generation from nonproductive coal by addition of nutrients or a microbial consortium. *Applied and environmental microbiology* **76**: 7013-7022.
Jones MB, Highlander SK, Anderson EL*, et al.* (2015) Library preparation methodology can influence genomic and functional predictions in human microbiome research. *Proceedings of the National Academy of Sciences* **112**: 14024.
Katayama T, Yoshioka H, Muramoto Y, Usami J, Fujiwara K, Yoshida S, Kamagata Y & Sakata S (2015) Physicochemical impacts associated with natural gas development on methanogenesis in deep sand aquifers. *The ISME Journal* **9**: 436-446.
Kirk M, Wilson B, Marquart K, Zeglin L, Vinson D & Flynn T (2015) Solute concentrations influence microbial methanogenesis in coal-bearing strata of the Cherokee basin, USA. *Frontiers in Microbiology* **6**.
Kleinsteuber S, Schleinitz KM & Vogt C (2012) Key players and team play: anaerobic microbial communities in hydrocarbon-contaminated aquifers. *Applied Microbiology and Biotechnology* **94**: 851-873.
Kurth JM, Op den Camp HJM & Welte CU (2020) Several ways one goal—methanogenesis from unconventional substrates. *Applied Microbiology and Biotechnology* **104**: 6839-6854.

838 Laczny CC, Sternal T, Plugaru V, Gawron P, Atashpendar A, Margossian HH,
839 Coronado S, der Maaten Lv, Vlassis N & Wilmes P (2015) VizBin - an application
840 for reference-independent visualization and human-augmented binning of
841 metagenomic data. *Microbiome* **3**: 1.
842 Lai M-C, Lin C-C, Yu P-H, Huang Y-F & Chen S-C (2004) Methanocalculus
843 chunghsingensis sp. nov., isolated from an estuary and a marine fishpond in
844 Taiwan. *International Journal of Systematic and Evolutionary Microbiology* **54**:
845 183-189.
846 Lai M-C, Chen S-C, Shu C-M, Chiou M-S, Wang C-C, Chuang M-J, Hong T-Y,
847 Liu C-C, Lai L-J & Hua JJ (2002) Methanocalculus taiwanensis sp. nov., isolated
848 from an estuarine environment. *International Journal of Systematic and*
849 *Evolutionary Microbiology* **52**: 1799-1806.
850 Lawson CE, Strachan CR, Williams DD, Koziel S, Hallam SJ, Budwill K & Liu S-J
851 (2015) Patterns of Endemism and Habitat Selection in Coalbed Microbial
852 Communities. *Applied and Environmental Microbiology* **81**: 7924-7937.
853 Liu Y & Whitman WB (2008) Metabolic, phylogenetic, and ecological diversity of
854 the methanogenic archaea. *Ann N Y Acad Sci* **1125**: 171-189.
855 Lozupone CA & Knight R (2007) Global patterns in bacterial diversity.
856 *Proceedings of the National Academy of Sciences* **104**: 11436.
857 Luton PE, Wayne JM, Sharp RJ & Riley PW (2002) The mcrA gene as an
858 alternative to 16S rRNA in the phylogenetic analysis of methanogen populations
859 in landfill. *Microbiology* **148**: 3521-3530.
860 Margesin R & Schinner F (2001) Biodegradation and bioremediation of
861 hydrocarbons in extreme environments. *Appl Microbiol Biotechnol* **56**: 650-663.
862 Mayumi D, Mochimaru H, Tamaki H, Yamamoto K, Yoshioka H, Suzuki Y,
863 Kamagata Y & Sakata S (2016) Methane production from coal by a single
864 methanogen. *Science* **354**: 222-225.
865 McKay LJ, Smith HJ, Barnhart EP, Schweitzer HD, Malmstrom RR, Goudeau D
866 & Fields MW (2022) Activity-based, genome-resolved metagenomics uncovers
867 key populations and pathways involved in subsurface conversions of coal to
868 methane. *The ISME Journal* **16**: 915-926.
869 Menzel P, Ng KL & Krogh A (2016) Fast and sensitive taxonomic classification
870 for metagenomics with Kaiju. *Nature Communications* **7**: 11257.
871 Meslé M, Dromart G & Oger P (2013) Microbial methanogenesis in subsurface oil
872 and coal. *Research in Microbiology* **164**: 959-972.
873 Milici RC & Polyak DE (2014) Coalbed-methane production in the Appalachian
874 basin. (Ruppert LF & Ryder RT, eds.), p.^pp. 29. Reston, VA.
875 Miller CS, Baker BJ, Thomas BC, Singer SW & Banfield JF (2011) EMIRGE:
876 reconstruction of full-length ribosomal genes from microbial community short
877 read sequencing data. *Genome Biology* **12**: R44.
878 Mori K, Yamamoto H, Kamagata Y, Hatsu M & Takamizawa K (2000)
879 Methanocalculus pumilus sp. nov., a heavy-metal-tolerant methanogen isolated
880 from a waste-disposal site. *International Journal of Systematic and Evolutionary*
881 *Microbiology* **50**: 1723-1729.

Morrissey EM & Franklin RB (2015) Evolutionary history influences the salinity preference of bacterial taxa in wetland soils. *Frontiers in microbiology* **6**: 1013-1013.
Nurk S, Meleshko D, Korobeynikov A & Pevzner PA (2017) metaSPAdes: a new versatile metagenomic assembler. *Genome Res* **27**: 824-834.
Ollivier B, Fardeau M-L, Cayol J-L, Magot M, Patel BKC, Prensier G & Garcia J-L (1998) Methanocalculus halotolerans gen. nov., sp. nov., isolated from an oil-producing well. *International Journal of Systematic and Evolutionary Microbiology* **48**: 821-828.
Orem WH, Voytek MA, Jones EJ, Lerch HE, Bates AL, Corum MD, Warwick PD & Clark AC (2010) Organic intermediates in the anaerobic biodegradation of coal to methane under laboratory conditions. *Organic Geochemistry* **41**: 997-1000.
Oren A (2008) Microbial life at high salt concentrations: phylogenetic and metabolic diversity. *Saline Syst* **4**: 2.
Parks DH, Imelfort M, Skennerton CT, Hugenholtz P & Tyson GW (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* **25**: 1043-1055.
Penner TJ, Foght JM & Budwill K (2010) Microbial diversity of western Canadian subsurface coal beds and methanogenic coal enrichment cultures. *International Journal of Coal Geology* **82**: 81-93.
Rajaram V DS, and Parameswaran K (2005) *Sustainable Mining Practices-A Global Perspective*. CRC Press, London.
Ritter DJ, Vinson DS, Barnhart EP, Akob DM, Fields MW, Cunningham AB, Orem WH & McIntosh JC (2015) Enhanced microbial coalbed methane generation: A review of research, commercial activity, and remaining challenges. *International Journal of Coal Geology* **146**: 28-41.
Robbins SJ, Evans PN, Parks DH, Golding SD & Tyson GW (2016) Genome-Centric Analysis of Microbial Populations Enriched by Hydraulic Fracture Fluid Additives in a Coal Bed Methane Production Well. *Frontiers in Microbiology* **7**.
Rodriguez-R LM & Konstantinidis KT (2016) The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. *PeerJ Preprints* **4**: e1900v1901.
Schweitzer H, Ritter D, McIntosh J, Barnhart E, Cunningham AB, Vinson D, Orem W & Fields MW (2019) Changes in microbial communities and associated water and gas geochemistry across a sulfate gradient in coal beds: Powder River Basin, USA. *Geochimica et Cosmochimica Acta* **245**: 495-513.
Schweitzer HD, Smith HJ, Barnhart EP, McKay LJ, Gerlach R, Cunningham AB, Malmstrom RR, Goudeau D & Fields MW (2022) Subsurface hydrocarbon degradation strategies in low- and high-sulfate coal seam communities identified with activity-based metagenomics. *npj Biofilms and Microbiomes* **8**: 7.
Seemann T (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**: 2068-2069.
Shaffer M, Borton MA, McGivern BB*, et al.* (2020) DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Research* **48**: 8883-8900.

927 Shelton JL, Akob DM, McIntosh JC, Fierer N, Spear JR, Warwick PD & McCray
928 JE (2016) Environmental Drivers of Differences in Microbial Community Structure
929 in Crude Oil Reservoirs across a Methanogenic Gradient. *Frontiers in*
930 *Microbiology* **7**.
931 Sorokin DY, Abbas B, Merkel AY, Rijpstra WIC, Damsté JSS, Sukhacheva MV &
932 van Loosdrecht MCM (2015) Methanosalsum natronophilum sp. nov., and
933 Methanocalculus alkaliphilus sp. nov., haloalkaliphilic methanogens from
934 hypersaline soda lakes. *International Journal of Systematic and Evolutionary*
935 *Microbiology* **65**: 3739-3745.
936 Strąpoć D, Mastalerz M, Dawson K, Macalady J, Callaghan AV, Wawrik B, Turich
937 C & Ashby M (2011) Biogeochemistry of Microbial Coal-Bed Methane. *Annual*
938 *Review of Earth and Planetary Sciences* **39**: 617-656.
939 Strąpoć D, Picardal Flynn W, Turich C*, et al.* (2008) Methane-Producing
940 Microbial Community in a Coal Bed of the Illinois Basin. *Applied and*
941 *Environmental Microbiology* **74**: 2424-2432.
942 Susilawati R, Evans PN, Esterle JS, Robbins SJ, Tyson GW, Golding SD &
943 Mares TE (2015) Temporal changes in microbial community composition during
944 culture enrichment experiments with Indonesian coals. *International Journal of*
945 *Coal Geology* **137**: 66-76.
946 Tamura K, Stecher G, Peterson D, Filipski A & Kumar S (2013) MEGA6:
947 Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol* **30**: 2725-
948 2729.
949 Team RC (2018) R: A language and environment for statistical computing. *R*
950 *Foundation for Statistical Computing, Vienna, Austria*.
951 Thielemann T, Cramer B & Schippers A (2004) Coalbed methane in the Ruhr
952 Basin, Germany: a renewable energy resource? *Organic Geochemistry* **35**: 1537-
953 1549.
954 Tierney M & Young LY (2010) Anaerobic Degradation of Aromatic Hydrocarbons.
955 *Handbook of Hydrocarbon and Lipid Microbiology,*(Timmis KN, ed.) p.^pp. 925-
956 934. Springer Berlin Heidelberg, Berlin, Heidelberg.
957 Van Voast WA (2003) Geochemical signature of formation waters associated
958 with coalbed methane. *AAPG Bulletin* **87**: 667-676.
959 Vick SHW, Greenfield P, Tetu SG, Midgley DJ & Paulsen IT (2019) Genomic and
960 phenotypic insights point to diverse ecological strategies by facultative
961 anaerobes obtained from subsurface coal seams. *Scientific Reports* **9**: 16186.
962 Vick SHW, Greenfield P, Pinetown KL, Sherwood N, Gong S, Tetu SG, Midgley
963 DJ & Paulsen IT (2019) Succession Patterns and Physical Niche Partitioning in
964 Microbial Communities from Subsurface Coal Seams. *iScience* **12**: 152-167.
965 Waldron PJ, Petsch ST, Martini AM & Nüsslein K (2007) Salinity constraints on
966 subsurface archaeal diversity and methanogenesis in sedimentary rock rich in
967 organic matter. *Appl Environ Microbiol* **73**: 4171-4179.
968 Widdel F & Grundmann O (2010) Biochemistry of the Anaerobic Degradation of
969 Non-Methane Alkanes. *Handbook of Hydrocarbon and Lipid*
970 *Microbiology,*(Timmis KN, ed.) p.^pp. 909-924. Springer Berlin Heidelberg, Berlin,
971 Heidelberg.

Widdel F & Musat F (2010) Diversity and Common Principles in Enzymatic Activation of Hydrocarbons. *Handbook of Hydrocarbon and Lipid Microbiology,*(Timmis KN, ed.) p.^pp. 981-1009. Springer Berlin Heidelberg, Berlin, Heidelberg.

Yamamoto K, Tamaki H, Cadillo-Quiroz H, Imachi H, Kyrpides N, Woyke T, Goodwin L, Zinder SH, Kamagata Y & Liu W-T (2014) Complete Genome Sequence of Methanoregula formicica SMSPT, a Mesophilic Hydrogenotrophic Methanogen Isolated from a Methanogenic Upflow Anaerobic Sludge Blanket Reactor. *Genome announcements* **2**: e00870-00814.

Yashiro Y, Sakai S, Ehara M, Miyazaki M, Yamaguchi T & Imachi H (2011) Methanoregula formicica sp. nov., a methane-producing archaeon isolated from methanogenic sludge. *Int J Syst Evol Microbiol* **61**: 53-59.

Zhang J, Liang Y, Pandey R & Harpalani S (2015) Characterizing microbial communities dedicated for conversion of coal to methane in situ and ex situ. *International Journal of Coal Geology* **146**: 145-154.

Zhilina TN, Zavarzina DG, Kevbrin VV & Kolganova TV (2013) Methanocalculus natronophilus sp. nov., a new alkaliphilic hydrogenotrophic methanogenic archaeon from a soda lake, and proposal of the new family Methanocalculaceae. *Microbiology* **82**: 698-706.

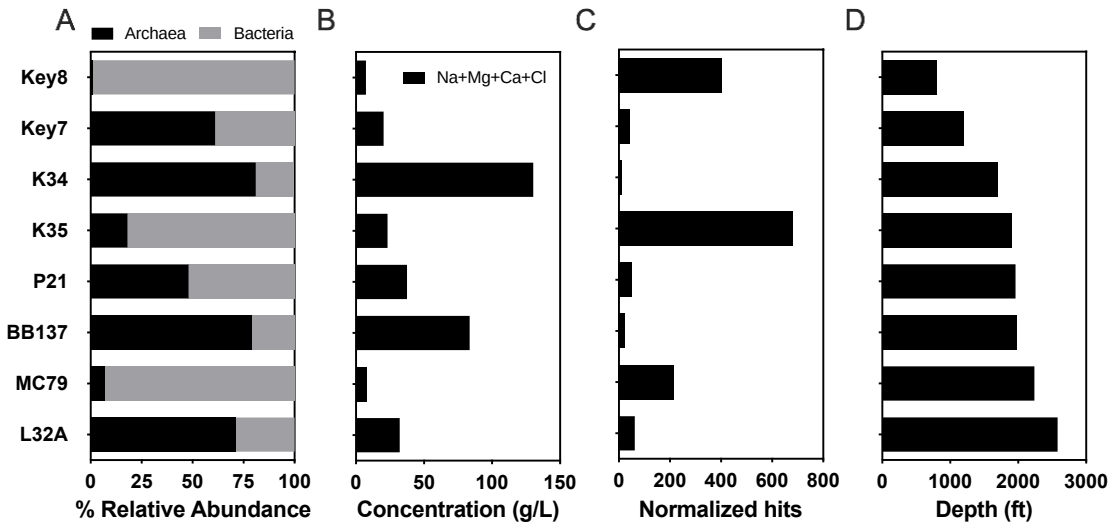


Figure 1. Read-based taxonomy profiling, and geochemistry of eight Appalachian Basin produced water samples. A) Kingdom-level taxonomy of coal seam metagenomes from unassembled sequencing reads. B) Major ions found in produced water samples. C) Normalized number of reads mapping to hydrocarbon degradation pathways. D) Depth profile of produced water samples. A positive correlation was observed between Archaea relative abundance (A) and salinity (B) within each sample, and a negative correlation was observed between Bacteria relative abundance (A) and salinity (B) (Supplementary Table S1).

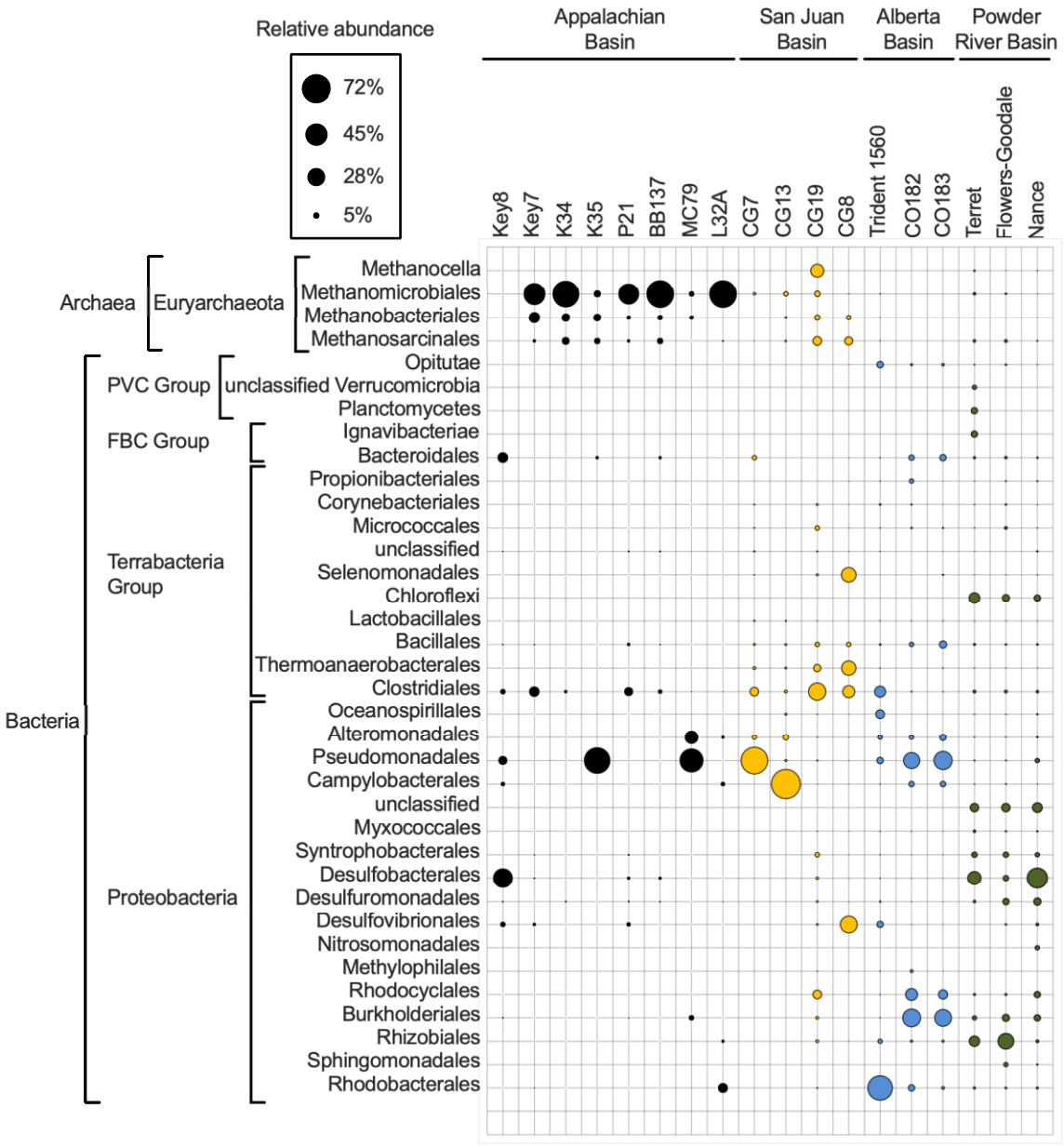
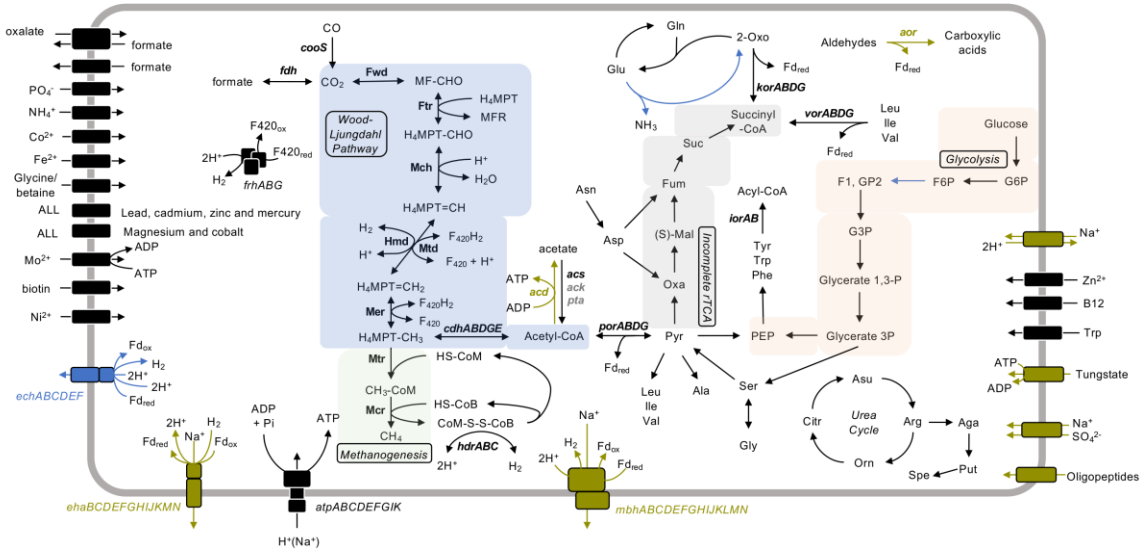


Figure 2. Taxonomic comparison of coalbed methane metagenomes. Order-level taxonomy of eight Appalachian Basin metagenomes were compared to coalbed methane metagenomes from the Alberta Basin, Powder River Basin, and San Juan Basin. The relative abundances of each order were determined with unassembled metagenome reads using Kaiju. Relative abundances were calculated based upon the percentage of classified reads for each sample.

Blue =Methanoregula, Yellow – Methanocalculus, Black =both, Grey =neither



1011

1012 **Figure 3. Select metabolic pathways found in the Methanomicrobiales MAGs.**

1013 Genes/pathways found in *Methanocalculus* MAGs (yellow), *Methanoregula* MAGs

1014 (blue), both (black), or neither (grey). Details for genes and corresponding proteins

1015 they encode can be found in Supplementary File 2.

1016

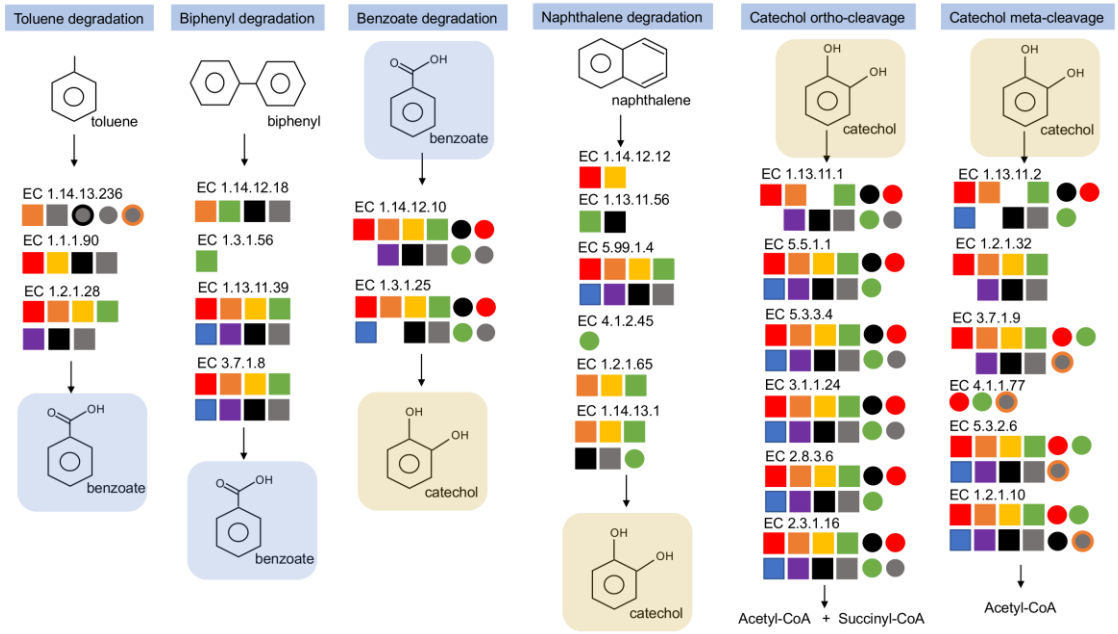


Figure 4. Overview of select hydrocarbon degradation pathways encoded in Appalachian Basin metagenomes/MAGs. Unassembled metagenome reads (squares), or metagenome-assembled genomes (circles) were examined for the presence of hydrocarbon degradation pathways, including toluene, biphenyl, benzoate, and naphthalene degradation. Catechol meta- and ortho-cleavage pathways were also examined. Metagenome samples are denoted as follows: Key8 (red), Key7 (orange), K34 (yellow), K35 (green), P21 (blue), BB137 (violet), MC79 (black), and L32A (grey). MAGs are denoted as follows: *Pseudomonas* MC79 (black), *Pseudomonas* Key8 (red), *Pseudomonas* K35 (green), *Marinobacter* L32A (grey/orange), *Hyphomonas* L32A (grey/black), and *Roseovarius* L32A (gray).